The Latent Doctor Model for Modeling Inter-Observer Variability

Jasper Linmans, Emiel Hoogeboom, Jeroen van der Laak, and Geert Litjens

***Abstract*—Many inherently ambiguous tasks in medical imaging suffer from inter-observer variability, resulting in a reference standard defined by a distribution of labels with high variance. Training only on a consensus or majority vote label, as is common in medical imaging, discards valuable information on uncertainty amongst a panel of experts. In this work, we propose to train on the full label distribution to predict the uncertainty within a panel of experts and the most likely ground-truth label. To do so, we propose a new stochastic classification framework based on the conditional variational auto-encoder, which we refer to as the Latent Doctor Model (LDM). In an extensive comparative analysis, we compare the LDM with a model trained on the majority vote label and other methods capable of learning a distribution of labels. We show that the LDM is able to reproduce the reference-standard distribution significantly better than the majority vote baseline. Compared to the other baseline methods, we demonstrate that the LDM performs best at modeling the label distribution and its corresponding uncertainty in two prostate tumor grading tasks. Furthermore, we show competitive performance of the LDM with the more computationally demanding deep ensembles on a tumor budding classification task.**

Manuscript received 17 November 2022; revised 15 May 2023, 21 July 2023, and 28 August 2023; accepted 8 October 2023. Date of publication 13 October 2023; date of current version 5 January 2024. This work was supported by European Union's Horizon 2020 Research and Innovation Program and EFPIA. The work of Jeroen van der Laak and Geert Litjens was supported by Innovative Medicines Initiative 2 Joint Undertaking under Grant 945358. The work of Geert Litjens was supported in part by Dutch Cancer Society (KWF) under Grant KUN 2015-7970, and in part by Dutch Research Council (NWO) under Grant 91618152. *(Corresponding author: Jasper Linmans.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by RadboudUMC under Application No. 2016-2275, and performed in line with the (Name of Specific Declaration).

Jasper Linmans is with the Computational Pathology Group, Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, 6525GA Nijmegen, The Netherlands (e-mail: jasper.linmans@radboudumc.nl).

Emiel Hoogeboom is with the Independent researcher, 1082MM Amsterdam, The Netherlands (e-mail: e.hoogeboom@outlook.com).

Jeroen van der Laak is with the Computational Pathology Group, Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, 6525GA Nijmegen, The Netherlands, and with the Center for Medical Image Science and Visualization, Linköping University, SE-58183 Linköping, Sweden, and also with the Chief Scientific Officer (CSO) and shareholder of Aiosyn BV Nijmegen, The Netherlands (e-mail: jeroen.vanderlaak@radboudumc.nl).

Geert Litjens is with the Computational Pathology Group, Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, 6525GA Nijmegen, The Netherlands, and also with the shareholder of Aiosyn BV, The Netherlands (e-mail: geert.litjens@radboudumc.nl).

Digital Object Identifier 10.1109/JBHI.2023.3323582

***Index Terms*—Digital pathology, stochastic classification, latent variable models, uncertainty estimation.**

I. Introduction

MANY diagnostic tasks in medical imaging suffer from high levels of inter-observer variability. For instance, tumor grading in prostate biopsies following the Gleason grading standard [1] involves inherent ambiguities, such as tumors whose histology is on the border between Gleason patterns [2], [3], [4]. Especially for complex cases, even experienced pathologists do not show good inter-observer agreement [2], [3], [4], [5]. Disagreement between experts results in a distribution of labels, complicating training and evaluation of conventional deep learning methods because a reliable reference standard is lacking.

Recent work on applying deep learning to automate prostate tumor detection mitigates these issues by evaluating on a consensus reference standard based on multiple rounds of re-grading [6], [7]. A more basic yet popular approach determines the reference standard based on simple majority voting. However, reducing the distribution of labels to a single ground truth discards any information about the disagreement amongst the panel of experts, which could indicate the difficulty of the case [5]. Therefore, modeling the full label distribution of experts might be equally important as modeling the most likely label. When trained correctly, the spread of the predicted distribution can be used as a surrogate for inter-observer variability. If the spread of the predicted distribution correlates with the inter-observer variability between experts, highly ambiguous cases can be discriminated from more straightforward cases based on the model output. After training, the most ambiguous cases could then be flagged and removed from an automated diagnostic pipeline, followed by further diagnostic tests or expert supervision.

The main goal of this work is to effectively predict the distribution of labels of a panel of experts instead of just the most likely label. To do so, we introduce a stochastic classification framework that provides multiple classification hypotheses for ambiguous images using a single classifier. The proposed framework combines a conditional variational auto-encoder (CVAE) [8] with a DeepSet encoder [9] that can capture annotator variability in a low-dimensional latent space. After training, we evaluate the ability to replicate the reference-standard label distribution and the corresponding disagreement between experts on different classification tasks in medical imaging.

A body of work with different approaches towards dealing with inter-observer variability in deep learning exists. A recent review evaluates relevant methods in medical imaging [10]. This includes a method used in chest X-rays based on label smoothing

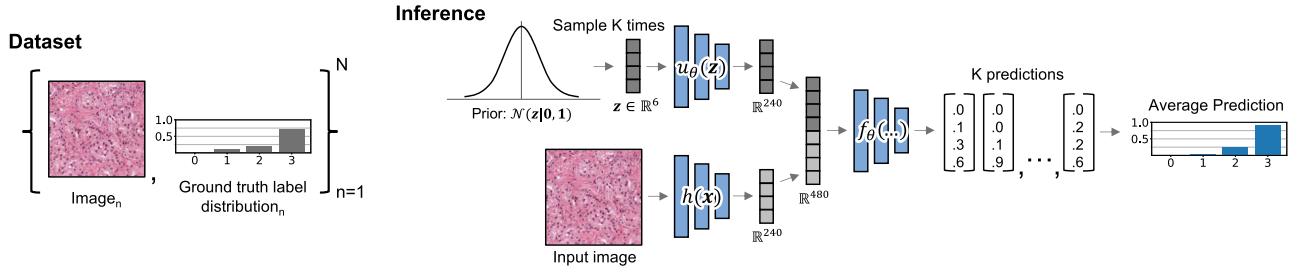

Fig. 1. Graphical representation of a dataset relevant for this work with the objective of learning a distribution of labels, and the steps involved during inference of the proposed Latent Doctor Model (LDM). We assume each case in the test set to be labeled by multiple doctors. The collection of one-hot labels can be illustrated with a histogram, referred to as the ground truth label distribution. During inference, the classifier of the LDM $f_\theta(y|\mathbf{x},\mathbf{z})$ is conditioned on samples from the prior Gaussian distribution, resulting in a stochastic classifier to model the predictive distribution $p_\theta(y|\mathbf{x})$. We adopt a pre-trained feature extractor $h(\mathbf{x})$ and a fully connected layer $u_\theta(\mathbf{z})$ to upsample $\mathbf{z}$, before feeding the data to the classifier. Inference for the Conditional LDM is similar, but the standard Gaussian prior is replaced with a learnable conditional prior distribution $p_\theta(\mathbf{z}|\mathbf{x})$.

to prevent overconfident predictions on training samples that contain mislabeled data [11]. Prior work on noisy labels captures the label noise of individual annotators [12]. However, their method assumes that the label noise is independent from the input image [12]. In contrast, our proposed method and corresponding baseline models are conditioned on the input image to capture more complex relationships. Related work improves classification accuracy by re-weighting the loss function for data samples with likely incorrect labels in both chest X-rays [13] and whole-slide digital pathology images [14]. However, these methods consider a setting of noisy labels, assuming some of the annotations to be incorrect, and focus on correcting annotations to train on cleaner data or model the trustworthiness of annotators. In contrast to our work, these works see inter-observer variability as a factor that needs to be reduced instead of an additional source of information relevant to modeling the inherent ambiguities of certain diagnostic tasks.

Directly modeling the distribution of labels in the context of classification remains relatively unexplored, especially in medical imaging. The works most similar to ours, model a distribution of *segmentation masks* in thoracic CT, lung CT, and prostate MR data [15], [16], [17]. To do so, they propose a probabilistic framework based on a combination of the CVAE and a U-Net architecture [18] to successfully model a diverse set of plausible segmentation masks in medical imaging. In essence, our proposed stochastic classification framework belongs to the same family as this probabilistic U-Net [15], but then for the context of *classification*.

The main contributions of this work are: (1) We develop a novel stochastic classification framework based on a CVAE with a DeepSet encoder and a stochastic classifier to model a distribution of class labels. We refer to the proposed framework as the Latent Doctor Model (LDM). (2) We propose the LDM with a prior based on a simple Gaussian with zero mean and unit variance, as well as a conditional model (CLDM), where the prior is modeled by a neural network conditioned on the input image. (3) We propose to compare against deep ensembles and a method based on knowledge distillation and quantitatively compare the results on three datasets in medical imaging. In contrast to the work which is most similar to ours [15], we compare against a more diverse ensemble baseline with members trained on individual experts instead of a regularly trained ensemble based on random initialization [19]. (4) Demonstrating the importance of modeling the full label distribution, we show that the different methods can reproduce the reference-standard distribution and its corresponding inter-observer variability significantly better than the majority vote baseline. Specifically, we show that the LDM is best at capturing the label distribution in two of the three datasets.

II. Preliminaries and Problem Statement

Throughout this work, we consider a classification setting with labels provided by multiple doctors $d \in \{1, 2, \ldots, D\}$. Let a dataset be noted as $\mathcal{D}_{train} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, where $\mathbf{x}_n$ represents one of N data points and $\mathbf{y}_n$ the label distribution determined by the set of annotations available for the n'th data point. Here, $\mathbf{y}_n$ is a normalized vector of individual annotations y_n^d; representing the histogram of class counts. We refer to this distribution of labels as the reference-standard or the ground-truth distribution. See Fig. 1 for a graphical depiction of such a dataset. Throughout the experiments, we will consider two types of ground-truth label distributions: the dense label scenario in which each image is annotated by every doctor d such that $\mathcal{D}_{train} = \{(\mathbf{x}_n, y_n^1, \ldots, y_n^D)\}_{n=1}^N$, as well as the limited label scenario where each image is annotated by only one, or a subset of doctors.

The task is to learn a classifier $f(\mathbf{x})$, which takes in an image and predicts the label distribution for $\mathbf{y}$. In this work, we present a novel stochastic classification framework based on a conditional variational auto-encoder (CVAE) [8], to model the probability distribution of labels given an image, enabling a one-to-many mapping.

A. The Conditional Variational Auto-Encoder

Recent work on variational auto-encoders [20] presents the conditional variational auto-encoder [8], to approximate a distribution $p_\theta(\mathbf{y}|\mathbf{x})$. The conditional generative process is as follows: for a given observation $\mathbf{x}$, $\mathbf{z}$ is drawn from a prior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ and the distribution of $\mathbf{y}$ is conditioned on both the input and the latent vector. The resulting predictive distribution is defined as:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z}) p_\theta(\mathbf{z}|\mathbf{x}) d\mathbf{z}. \quad (1)$$

In a classification or segmentation setting, $\mathbf{z} \in \mathbb{R}^m$ defines a continuous latent variable capturing annotator variability such that $p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})$ captures multiple plausible classification hypotheses. Due to the intractable posterior distribution $p(\mathbf{z}|\mathbf{x},\mathbf{y})$,

the predictive distribution $p_\theta(\mathbf{y}|\mathbf{x})$ can not be optimized directly. Instead, the predictive distribution can be optimized through the evidence lower bound (ELBO), similarly as for the VAE [20]:

$$\begin{aligned}
&\log p_\theta(\mathbf{y}|\mathbf{x}) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{y},\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\right] + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\|p(\mathbf{z}|\mathbf{x},\mathbf{y})\right) \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})\|p_\theta(\mathbf{z}|\mathbf{x})\right)
\end{aligned} \tag{2}$$

where $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ is the variational posterior parametrized by variational parameters ϕ. The second right-hand side term of the ELBO defines the Kullback-Leibler divergence between the prior $p_\theta(\mathbf{z}|\mathbf{x})$ and the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, which is sometimes combined with the first term as a weighted sum with a weighting factor β, as proposed by [21].

Both the prior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ as well as the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ are modeled using Gaussian distributions with a diagonal covariance matrix whose parameters are the output of a neural network. As a result, the CVAE uses three trainable networks [15], [16], including the classification network for $p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})$, while the VAE only uses two. However, as argued by the original authors [8], the conditioning of $\mathbf{z}$ can be easily relaxed to make the latent variables statistically independent of input variables such that $p_\theta(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})$, with $p(\mathbf{z})$ a standard Gaussian distribution. When presenting the CVAE used for classification in the next section, we first start with using p($\mathbf{z}$), a standard Gaussian prior distribution. Afterwards, we will define a conditional variant of our method by re-introduce the conditioning of the prior distribution on the input.

B. Lacking Information in the Context of Classification

CVAEs have been successfully applied to produce a diverse set of plausible predictions in the context of *segmentation* in medical imaging [15], [16]. By combining the CVAE with a U-Net architecture [18], the probabilistic U-Net is capable of modeling a distribution of segmentation masks $\mathbf{y}$. To do so, an encoder network, representing the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$, is required to capture annotator variability in the latent space. Here, the high-dimensionality of a ground truth segmentation mask $\mathbf{y}$ enables the encoding of small variations between annotators. In contrast, the application of CVAE in the context of *classification* is impeded by the low-dimensionality of a single class label to model the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x},y)$. Specifically, a single (one-hot) class label y only contains a limited amount of information for the encoder to capture annotator variability.

To overcome the problem related to the amount of information available, we propose to train an encoder using a collection of image-label pairs, in contrast to a single data point, as is the case in the probabilistic U-Net. Doing so, our approach is fundamentally different from the probabilistic U-Net in that we adopt a CVAE to learn a regularised latent representation of doctors to model a univariate variable y. In contrast, the probabilistic U-Net learns a joint distribution of all pixels in a segmentation map to model a consistent interpretation of a whole image.

III. The Latent Doctor Model

We introduce the Latent Doctor Model (LDM), a probabilistic classification model based on the CVAE, that is capable of efficiently producing multiple plausible classification hypotheses. Here, a latent space encodes the possible label variants.

To do so, we introduce a set $\mathcal{S}^d$ with image-label pairs sampled from the train set. Without changing the train set itself, we define $\mathcal{S}^d$ to contain M individual samples $(\mathbf{x}_m, y_m^d)$, such that $\mathcal{S}^d \sim \mathcal{D}_{train}(\mathcal{S}^d)$. Using a standard Gaussian prior $p(\mathbf{z})$, we rewrite the lower bound of the CVAE (2) into:

$$\begin{aligned}
\log p_\theta(y|\mathbf{x},\mathcal{S}^d) \geq\ & \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{S}^d)}\left[\log p_\theta(y|\mathbf{x},\mathbf{z})\right] \\
& - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathcal{S}^d)\middle\|p(\mathbf{z})\right)
\end{aligned} \tag{3}$$

where the posterior $q_\phi(\mathbf{z}|\mathcal{S}^d)$ is conditioned on the collection of image-label pairs. Similar to the probabilistic U-Net, we model the posterior $q_\phi(\mathbf{z}|\mathcal{S}^d)$ with a neural network which maps the input to a position $\boldsymbol{\mu}_{post}(\mathcal{S}^d)$, with some variance $\boldsymbol{\sigma}_{post}(\mathcal{S}^d)$. Optimizing the lower bound (3), therefore minimizing the KL divergence between the posterior $q_\phi(\mathbf{z}|\mathcal{S}^d)$ and the prior $p(\mathbf{z})$, will result in:

$$\mathbb{E}_{\mathcal{D}_{train}(\mathcal{S}^d)}[q_\phi(\mathbf{z}|\mathcal{S}^d)] \approx p(\mathbf{z}). \tag{4}$$

In other words, sampling from the prior distribution will resemble sampling from the trained posterior distribution and as a result, $\mathcal{S}^d$ is only required during training. Instead, during inference, we sample $\mathbf{z}$ directly from the standard Gaussian prior $p(\mathbf{z})$, as is normal for the (C)VAE framework [8], [20], to model the predictive distribution:

$$p(y|\mathbf{x}) = \mathbb{E}_{p(\mathbf{z})}p(y|\mathbf{x},\mathbf{z}) \tag{5}$$

which can be approximated using some finite number of samples from the prior $p(\mathbf{z})$. See Fig. 1 for the steps involved during inference.

A. Training Objective of the LDM

Optimizing (3) based on training data points $(\mathbf{x}_n, y_n^d)$, will require a *classifier* $p_\theta(y|\mathbf{x},\mathbf{z})$ to predict the class label for annotator d, conditioned on a sample $\mathbf{z}$ from the posterior distribution. To enable the encoder $q_\phi(\mathbf{z}|\mathcal{S}^d)$ to capture annotator specific variability in this posterior distribution, we restrict $\mathcal{S}^d$ to only contain image-label pairs with labels from a shared doctor d. Specifically, we define $\mathcal{S}^d$ to be the set:

$$\mathcal{S}_n^d = \{(\mathbf{x}_m, y_m^d) \mid m \in I_d, m \neq n\} \tag{6}$$

of size M, with I_d the set of indices for which the d'th doctor provided a ground-truth label y_m^d.

In other words, $\mathcal{S}_n^d$ contains randomly sampled image-label pairs from the training set with labels from the d'th doctor, excluding the n'th data point. The constraint $m \neq n$, prevents the latent embedding from simply forwarding the ground-truth label y_n^d to the classifier during training. While sharing doctor index d between all labels in $\mathcal{S}_n^d$, enables $q_\phi(\mathbf{z}|\mathcal{S}^d)$ to capture doctor specific annotation variations. We refer to $\mathcal{S}_n^d$ as the *support set* for the image-label pair $(\mathbf{x}_n, y_n^d)$. To optimize (3), we define the following training objective for a single datapoint:

$$\begin{aligned}
\mathcal{L}(\mathbf{x},y,\mathcal{S}^d,\theta,\phi) = & \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{S}^d)}\left[\log p_\theta(y|\mathbf{x},\mathbf{z})\right] \\
& - \beta D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathcal{S}^d)\middle\|p(\mathbf{z})\right)
\end{aligned} \tag{7}$$

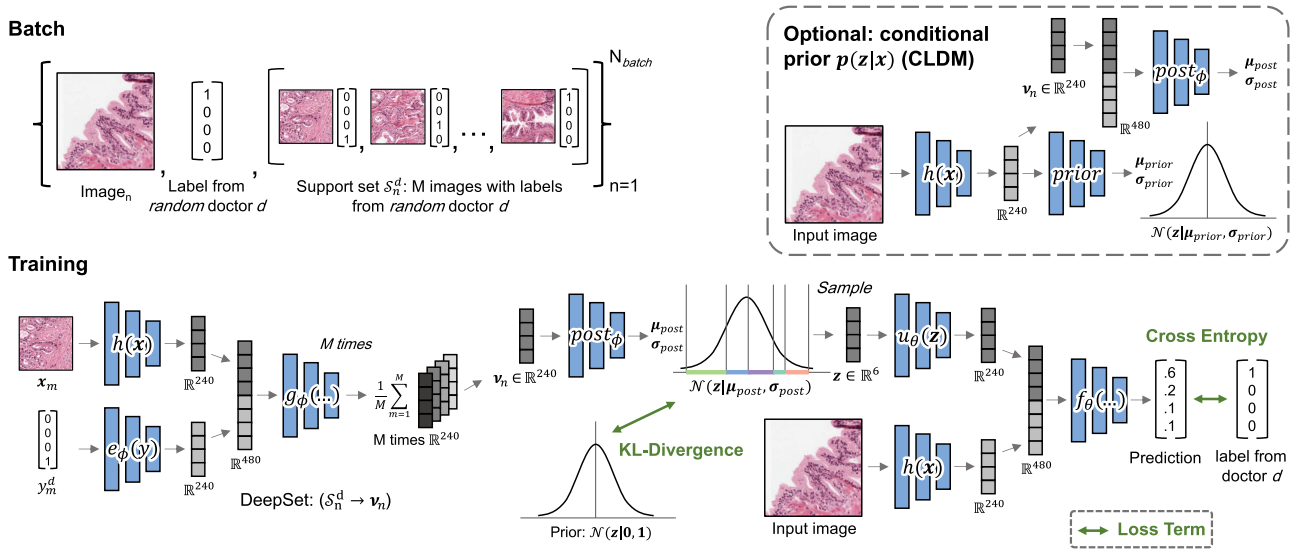

Fig. 2. Graphical representation of the training phase of the proposed (Conditional) Latent Doctor Model (LDM). In plain terms, we are asking a classifier $f_\theta(y|\mathbf{x},\mathbf{z})$ to predict what label a specific doctor d would have assigned to each case in the training set. To help the network make this decision, we provide examples of how the doctor scored other images in the support set $\mathcal{S}^d$. As illustrated, each entry within a mini-batch consists of an image, the associated label from a random doctor d and a corresponding support set. As a result, a *single classifier* is trained to map an input image to the ground-truth label of multiple doctors D. Here, the DeepSet summarizes the entire support set in a single vector $\boldsymbol{\nu}_n$, using a shared function g_ϕ with inputs from a feature extractor $h(\mathbf{x})$ and embedded labels $e_\phi(y)$. The posterior distribution is modeled by a simple feed forward network using $\boldsymbol{\nu}_n$ as input. Samples from the posterior distribution (upsampled by a fully connected layer $u_\theta(\mathbf{z})$) are used as input to the classifier along with a feature representations of the input $h(\mathbf{x})$. The top-right panel shows the training procedure of the CLDM, where the standard Gaussian prior is replaced with a learnable conditional prior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ using an additional prior network. To minimize the KL divergence between the prior and the posterior distribution, the image feature $h(\mathbf{x})$ is also fed to the posterior distribution. Loss terms are indicated in green.

with the corresponding training objective for the entire dataset defined as $\mathcal{L}(\theta,\phi) = \mathbb{E}_{x,y,d,\mathcal{S}^d\sim\mathcal{D}_{train}}[\mathcal{L}(\mathbf{x},y,\mathcal{S}^d,\theta,\phi)]$. Here, for each entry in a mini-batch, we first sample an image-label pair with a label from a random doctor index d. We then use index d to sample a support set $\mathcal{S}^d$ from the train set to define the approximate posterior distribution $q_\phi(\mathbf{z}|\mathcal{S}^d)$. See Fig. 2 for an illustration of a mini-batch and an overview for the steps involved during training. Similar to the probablistic U-Net, we combine the KL divergence term with the first term as a weighted sum with a weighting factor β [21].

B. The Conditional LDM

Similar to the probabilistic U-Net, we can condition the latent space of the LDM on the input image $\mathbf{x}$ in an effort to increase the model complexity of the latent space. Here the standard Gaussian prior distribution is replaced with a learned conditional distribution $p_\theta(\mathbf{z}|\mathbf{x})$, as discussed in section II-A. We refer to this method as the Conditional LDM (CLDM). This is done by introducing a prior network, in addition to the encoder and the classifier, that learns to map an input $\mathbf{x}$ to a position $\boldsymbol{\mu}_{prior}(\mathbf{x})$, with some variance $\boldsymbol{\sigma}_{prior}(\mathbf{x})$, similar to [15], [16]. The corresponding lower bound of the CLDM is as follows:

$$\log p_\theta(y|\mathbf{x},\mathcal{S}^d) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{S}^d)}\left[\log p_\theta(y|\mathbf{x},\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathcal{S}^d,\mathbf{x})\middle\|p_\theta(\mathbf{z}|\mathbf{x})\right) \quad (8)$$

where both the posterior $q_\phi(\mathbf{z}|\mathcal{S}^d,\mathbf{x})$ and the prior $p_\theta(\mathbf{z}|\mathbf{x})$ are conditioned on the input $\mathbf{x}$. See Fig. 2 for a graphical representation of the different networks defining both the LDM and the CLDM.

C. DeepSet Encoder

In contrast to the encoder of a regular CVAE, the encoder of the LDM encodes information from a set of data points. The structure of the set is random and does not need to be learned, instead the encoder should be invariant to permutations. To this end, we leverage DeepSets [9] to build permutation invariant representations of the support set $\mathcal{S}^d$. Here, each image $\mathbf{x}\in\mathbb{R}^I$ and corresponding one-hot label $y\in\mathbb{R}^C$ in the set, are transformed into some representation using a shared function. The resulting output vectors are then aggregated by simply taking the average, resulting in a single vector representation of the entire set:

$$\boldsymbol{\nu}_n = \frac{1}{M}\sum_{m=1}^{M} g_\phi(\mathbf{x}_m, y_m^d). \quad (9)$$

Here $g_\phi(\mathbf{x}_m, y_m^d)$ defines the representation of the m'th image-label pair in the support set $\mathcal{S}^d$.

This definition allows for any arbitrary shared function and may be defined by the user. Throughout this work, we model $g_\phi(\mathbf{x}_m, y_m^d)$ with a simple feed-forward network, which takes in a concatenation of a lower dimensional feature representation $h(\mathbf{x})$ and a learned embedding $e_\phi(y)$. For completeness, we define the feature extractor as: $h:\mathbb{R}^I\rightarrow\mathbb{R}^F$, and the learned embedding as $e_\phi:\mathbb{R}^C\rightarrow\mathbb{R}^F$ to map both inputs to a vector of similar length. As a result, we define the shared function as $g_\phi:\mathbb{R}^F\times\mathbb{R}^F\rightarrow\mathbb{R}^H$. See section IV-F for implementation details for each module, and Fig. 2 for an illustration of the DeepSet encoder framework used throughout this work.

The output vector $\boldsymbol{\nu}_n$ of the DeepSet model is then fed to a feed-forward neural network to produce the parameters of the approximate posterior distribution for the LDM:

$$q_\phi(\mathbf{z}|\mathcal{S}^d) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_{post}(\boldsymbol{\nu}_n), \boldsymbol{\sigma}_{post}(\boldsymbol{\nu}_n)). \tag{10}$$

The approximate posterior distribution for the CLDM is similar, but also conditioned on the input $\mathbf{x}$. Here, the parameters of the posterior distribution are defined as $\boldsymbol{\mu}_{post}(\boldsymbol{\nu}_n, h(\mathbf{x}))$ and $\boldsymbol{\sigma}_{post}(\boldsymbol{\nu}_n, h(\mathbf{x}))$. See Fig. 2 for an illustration. Note that, although we use DeepSet encoders throughout this work, the framework of the LDM is flexible towards other invariant functions on $\mathcal{S}^d$ as well.

D. Stochastic Classification

The classifier $p_\theta(y|\mathbf{x}, \mathbf{z})$ is modeled by a neural network which maps an input image and a sample from the posterior distribution $\mathbf{z} \in \mathbb{R}^E$, to a vector of size C, defined by the amount of classes in the dataset, such that:

$$\mathbf{y}^{pred} = f_\theta(\mathbf{x}_n, \mathbf{z}). \tag{11}$$

Again, this definition allows for any arbitrary function and may be defined by the user. Throughout this work, we model $f_\theta(\mathbf{x}_n, \mathbf{z})$ with a simple feed-forward network, which takes in a concatenation of a lower dimensional feature representation $h(\mathbf{x})$ and an upsampled latent vector $u_\theta(\mathbf{z})$. Here, we reuse the feature extractor $h(\mathbf{x})$ from the DeepSet encoder, and we define the upsampling layer as $u_\theta : \mathbb{R}^E \rightarrow \mathbb{R}^F$. The resulting classifier is defined by $f_\theta : \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}^C$. See Fig. 2 for an illustration of the steps involved in training the DeepSet encoder and the stochastic classifier end-to-end.

The output $\mathbf{y}^{pred}$ represents the vector of probabilities of the categorical distribution $Cat(\mathbf{y}|p = \mathbf{y}^{pred})$, which is standard for a classifier. During training, the log-likelihood of this categorical distribution is optimized through regular training using cross-entropy loss. Furthermore, the KL Divergence term of (7) can be optimized directly based on a closed form solution, since the posterior and the prior are modelled by Gaussian distributions in both the LDM and the CLDM. Both loss terms are indicated in green in Fig. 2.

Throughout this work, we use $\beta = 1\text{e}{-3}$ and a latent space with 6 dimensions to optimize (7) and its equivalent for the CLDM. During training, we sample support sets of size 16 or 32 to train the encoder. See section IV-F for further implementation details per experiment.

E. Inference

After training, the label distribution of the panel of experts can be modeled using multiple forward passes through the classifier $p_\theta(y|\mathbf{x}, \mathbf{z})$, conditioned on random samples $\mathbf{z}$, following (5). When trained correctly, different locations in the latent space encode different label variants (see section III-A), such that multiple forward passes through the network result in a distribution of predictions (11). To do so, $\mathbf{z}$ can be sampled directly from the prior distribution, eliminating the need for a support set $\mathcal{S}^d$ during inference, following (4). In other words, we can simply sample from a standard Gaussian distribution $p(\mathbf{z})$ for the LDM during inference. We refer back to Fig. 1 for an overview of the steps involved during inference. Note that we reuse the output of the feature extractor $h(\mathbf{x})$, when drawing samples from $p_\theta(y|\mathbf{x}, \mathbf{z})$, following (11). As a result, only the forward pass through the upsamling layer u_θ and the feed-forward network f_θ need to be recomputed several times. This results in a computationally efficient way to sample K predictions from the LDM to model the predictive distribution. Throughout this work, we use $K = 64$. See Algorithm 1 for a summary of the steps involved in training and inference.

Algorithm 1: The Latent Doctor Model. We Use $N \in [16, 32]$, $\beta = 1\text{e}{-3}$, and $K = 64$ Throughout The Experiments.

// Training phase
$\theta, \phi \leftarrow$ Initialize parameters
repeat
for *each entry* in *mini-batch* of size N **do**
  $\mathbf{x}_n \leftarrow$ Sample random image from $\mathcal{D}_{train}$;
  $d \leftarrow$ Sample doctor index from available annotations;
  $y_n^d \leftarrow$ Assign label from doctor d;
  $\mathcal{S}_n^d \leftarrow$ Sample support set, with labels from doctor d;
  $\boldsymbol{\nu}_n \leftarrow$ Forward pass $\mathcal{S}_n^d$ through DeepSet;
  $\mathbf{z}_n \sim \mathcal{N}(\mathbf{z} \mid \mu_{post}(\boldsymbol{\nu}_n), \sigma_{post}(\boldsymbol{\nu}_n))$ // Encoding;
  $\mathbf{y}_n^{pred} \leftarrow f_\theta(\mathbf{x}_n, \mathbf{z}_n)$ // Classification;
end for
Perform stochastic gradient descent using mini-batch to update parameters θ and ϕ:
$\nabla_{\theta,\phi} \frac{1}{N} \Sigma_{n=1}^N \log Cat(y_n^d|\mathbf{y}_n^{pred}) + \beta D_{\text{KL}}(q_\phi(\mathbf{z}|\mathcal{S}_n^d) \| p(\mathbf{z}))$
until convergence of parameters (θ, ϕ)
// Inference phase
$\mathbf{x}^* \leftarrow$ Unseen image;
$\mathbf{x}_{feat.}^* \leftarrow h(\mathbf{x}^*)$
for K steps **do**
  $\mathbf{z}_k \sim p(\mathbf{z})$
  $\mathbf{y}_k^{pred} \leftarrow f_\theta(\mathbf{x}_{feat.}^*, \mathbf{z}_k)$
end for
$\mathbf{y}^{pred} = \frac{1}{K} \sum_{k=1}^K \mathbf{y}_k^{pred}$ // Predictive distribution;

IV. Experimental Setup

We evaluate performance on three datasets, each with annotations provided by multiple annotators. This section first describes the datasets and corresponding training objectives (sections IV-A through IV-D). Afterward, the baseline methods (IV-E), implementation details (IV-F), and the evaluation metrics for correlating the disagreement between experts and the predictive disagreement (IV-G) will be described. All data used throughout this work has been part of prior publications. Here, most data was obtained from open source archives: IV-A, IV-B. For the material that was obtained in our hospital (IV-B and IV-D), the need for informed consent was waived by the local ethics review board (2016–2275).

A. Gleason Grading on Gleason2019

First, we use the publicly available training dataset from the Gleason2019 challenge [22], [23]. Detailed pixel-level annotations are provided by six pathologists on 244 prostate tissue micro-array (TMA) cores, based on four classes: benign and cancerous with Gleason patterns 3, 4, and 5. On average, every pathologist provides annotations for a subset of 197 cores, corresponding to the limited label scenario.

In this work, we follow the training and evaluation pipeline as proposed by [10], and train on patches with labels corresponding to the annotations of the central pixel. For each core, we extract 20 patches of size 279×279 randomly, and follow a 5-fold cross-validation framework. Each time, we train and tune a model on 80% of the TMA cores and their labels from the six pathologists, and then evaluate on the remaining 20% of the cores. During inference, we evaluate on the distribution of labels from the six pathologists and their consensus, modeled by the Simultaneous Truth and Performance Level Estimation (STAPLE) [24], as suggested by [10]. To improve robustness, we repeat each fold three times, resulting in a total of 15 independent train runs. We calculate and report the mean and standard deviation values for each performance metric across all 15 runs. Specifically, we measure the unweighted Cohen's kappa coefficient with the consensus label to enable comparisons with earlier work. Furthermore, we report the accuracy scores for distinguishing cancerous (Gleason patterns 3–5) from benign tissue, and separating high-grade (Gleason patterns 4 and 5) from low-grade (Gleason pattern 3) cancer, as reported by [10].

B. Slide-Level ISUP Grading

The second dataset we evaluate on, consists of 100 whole-slide images (WSIs) of prostate biopsies taken from the test set of [6]. Here, each WSI is annotated on a slide-level using the ISUP grading system which stratifies ratio's of Gleason patterns into five types, from 1 (low risk) to 5 (high risk) [25]. In total, the set of 100 WSIs are all annotated by 20 pathologists from 14 independent labs and ten countries, based on six classes (including the benign class). Similar to the training pipeline of the previous experiment, we repeat the training setup for 5 different folds. Each time, we train and tune models on approximately 80% of the set of WSIs and their labels from the 20 pathologists, and then evaluate on the remaining set of WSIs. Specifically, to deal with class imbalances we define each test set as a random subset of 18 WSIs, balanced across the six classes. Different from regular cross-validation, the resulting test sets can have overlapping data points.

For this task, we leverage techniques of weakly-supervised learning to summarize patch-level features into slide-level representations. Considering that Gleason patterns define morphological properties of epithelium tissue in particular, we train and evaluate only on patches extracted from the epithelial tissue. To do so, we first evaluate the epithelium segmentation model of [6], on all WSIs in the dataset. To artificially increase the amount of data points, as a way of data augmentation, we train on a randomly sampled subset of 128 epithelium patches of size 279×279 at a pixel spacing of $0.96\mu m$. During inference, we aggregate the predictions across all five folds and report the results on the total set of 90 predictions. To improve robustness, we repeat this process three times, and report the mean and standard deviation values for each metric across the three repetitions. To evaluate grading accuracy, we report the Cohen's kappa coefficient with a linear weighting.

To analyze the performance on ISUP grading in a *limited label scenario*, we repeat the experiment while removing some of the annotations in each training set. To do so, at the start of training we select a random subset of five annotations per WSI (fixed during training and the same for each method), effectively reducing the amount of annotations per doctor but not the total number of annotators. During inference, we evaluate on the test set using the full label distribution of 20 annotators.

C. Leveraging Pre-Trained Gleason Feature Extractors

To improve performance, for both the Gleason2019 and the ISUP grading tasks, we pre-train a Gleason pattern classifier on data from the PANDA challenge [26]. This feature extractor network $h(\mathbf{x})$ is previously introduced in section III-C, see Fig. 1, and Fig. 2. Specifically, we train a lightweight DenseNet architecture [27], as specified in [28], on 5060 WSIs and select patches in a 2:3 normal to tumor ratio of size 279×279 at a pixel spacing of $0.96\mu m$.

For the Gleason2019 dataset and before training, we finetune the pre-trained backbone model for each fold individually, using the TMA cores from the corresponding training sets and the consensus labels defined by the STAPLE method. Since the ISUP grading dataset originates from the same medical center as the set of 5060 WSIs, no finetuning is required. In both experiments, the pre-trained models are used to extract feature vectors of size 240 from the penultimate layer. We've experimented with training from scratch during preliminary analysis, for both tasks, but found that leveraging pre-trained models resulted in improved performance on the validation set. Additionally, the use of pre-trained feature extractors reduced the computational requirements by an order of magnitude. This allowed us to train computationally demanding ensemble baselines and repeat the experiments multiple times to estimate confidence intervals.

D. 3 k Buds

The third and final dataset that we use is in this work, is the 3000 bud study from [29]. Here, from a pool of 63 cytokeratin stained WSIs containing colorectal cancer cases with tumor budding, 3000 tumor bud candidates were selected. Each candidate bud is categorized by seven (of a total of nine) pathologists at a patch level as: either a tumor bud, a poorly differentiated cluster (PDC) or neither. In contrast to the previous two datasets, these labels are more categorical in nature. As such, modeling doctor specific annotation characteristics might be more difficult.

We split the 3000 buds in a train, valid, and test dataset on a slide-level and selected a random subset of 8 slides for both the validation and test splits. The remaining 47 slides were used for training. The resulting train, valid, and test datasets consisted of 2130, 152, and 718 bud candidates with patches extracted at a pixel spacing of $0.48\mu m$ of size 279×279. During evaluation, we report the Cohen's kappa coefficient (with a linear weighting) with the consensus label.

E. Baseline Methods

To compare the LDM with a regularly trained network, we include a baseline model trained on the *majority vote* label in each experiment.

1) Histogram of Doctors: Additionally, we train a model using the full label distribution. Here, the goal is to "distill" the knowledge from the panel of experts into a single model, similar to related work [30], [31], which transfers knowledge from a cumbersome teacher model to a smaller model. Since we train on the full histogram of doctor labels, we refer to this as the Histogram of Doctors (HoD) method. Without changing the majority vote model, we apply the cross-entropy loss using the

normalized reference-standard distribution $\mathbf{y}$:

$$\mathcal{H}(\mathbf{y}^{pred}, \mathbf{y}) = \sum_{c=1}^{C} y_c \log y_c^{pred} \tag{12}$$

with C the number of classes, $\mathbf{y}^{pred}$ the predicted distribution, and $\mathbf{y}$ the reference-standard label distribution. In contrast, in the case of the majority vote method, $\mathbf{y}$ is simply defined by a one-hot vector and the sum over C reduces to a single term.

2) Ensemble (Train One CNN Per Doctor): Finally, to model the set of annotators explicitly, we also include *deep ensembles* [19]. Here, each member of the ensemble is trained using the annotations and corresponding data points of a single annotator. During inference, the predictions of all individual members are averaged following:

$$\mathbf{y}^{pred} = \frac{1}{D} \sum_{d=1}^{D} \mathbf{y}_d^{pred} \tag{13}$$

with $\mathbf{y}_d^{pred}$ the prediction of each individual member. Compared to the LDM, this method requires significantly more computational resources to train and evaluate D individual members. For example, the cross-validation setup described previously, requires training and evaluating a total of 90 and 300 ensemble members in the Gleason2019 and ISUP grading experiments respectively.

We would like to emphasize that this ensemble is different from the originally proposed deep ensemble [19]. Here, we specifically try to capture the ground truth distribution of labels by training on all doctors individually, instead of training multiple randomly initialized ensemble members on the same majority vote label.

F. Implementation Details

This section summarizes the relevant details used to train and evaluate the different methods in each experiment.

1) Gleason Grading: Leveraging the pre-trained Gleason feature extractors $h(\mathbf{x})$, the training objective reduces to mapping feature vectors of size 240 to a class label. To this end, we implement the different classifiers $p_\theta(y|\mathbf{x})$ with a feed-forward neural network, consisting of three fully connected layers, reducing the amount of features with a factor of 0.5 at each layer. We use BatchNorm layers and ReLU non-linearities before and after each fully connected layer, and apply dropout ($p = 0.5$) after the second layer. We do not change this architecture for the classifier $p_\theta(y|\mathbf{x}, \mathbf{z})$, used in the LDM. However, we do increase the input dimensions to match the size of the image feature concatenated with a (upsampled) sample from the latent space $u_\theta(\mathbf{z})$. Here the upsampling layer u_θ is simply defined by a single fully connected layer, mapping the six dimensional latent vector to a vector of size 240. To train the DeepSet encoder $q_\phi(\mathbf{z}|\mathcal{S}^d)$, we repeat this architecture to define the shared DeepSet function g_ϕ, from (9). We define the size of the support set as 32 patches. The resulting DeepSet output $\boldsymbol{\nu}_n$ is then used as input to a fully connected layer with two output nodes, representing the mean and standard deviation values of the posterior distribution, following (10). The CLDM extends the LDM by adding a prior network which mimics the three-layer fully connected architecture, with only two output nodes: representing the mean and standard deviation values of the prior network, similar to (10). Finally, the CLDM also concatenates the feature vector $h(\mathbf{x})$ with $\boldsymbol{\nu}_n$ before feeding it to the posterior network, see Fig. 2.

2) ISUP Grading: To train the different ISUP classifiers and classify on a WSI-level (instead of a patch-level: Gleason grading), we add an additional attention layer to summarize a set of input features $h(\mathbf{x})$, using the framework of attention-based Multiple Instance Learning (MIL) [32]. Specifically, we consider each WSI datapoint as a bag of 128 features $h(\mathbf{x})$ of size 240, and use an attention layer to calculate its weighted average. The resulting vector of size 240 is then used for further processing, with the rest of the architecture identical to the previous experiment. To optimize the (C)LDM, we use a support set of size 16.

3) Tumor Budding: Last, we repeat the setup from the Gleason2019 experiment for the tumor budding dataset. Here we use a DenseNet backbone architecture, identical to the one used in the Gleason feature extractor (IV-C), to transform each input patch to a feature vector $h(\mathbf{x})$ of size 240, used as input for further processing by each classifier. Similarly, we use a support set of size 32 during training of the (C)LDM.

G. Evaluation Metrics

To enable comparison with prior work on these datasets, we evaluate the classification accuracy with task specific metrics, see sections IV-A through IV-D. More importantly, we also evaluate each method on the ability to capture the entire ground-truth label distribution and the corresponding inter-observer variability, using the following metrics.

1) KI-Divergence: Using the Kullback-Leibler divergence $D_{\mathrm{KL}}(\mathbf{y}\|\mathbf{y}^{pred})$ with $\mathbf{y}$ the ground-truth label distribution and $\mathbf{y}^{pred}$ the predicted distribution, we directly compare the two. Here, a lower value indicates a higher similarity between two distributions, with 0 the value corresponding to identity.

2) Spearman's Correlation Coefficient: To evaluate uncertainty, we measure the entropy of the predictive distribution, as is common in the uncertainty estimation literature. By evaluating Spearman's correlation coefficient between the entropy of the prediction and the entropy of the ground-truth distribution, we can measure how informative the predictive uncertainty is for the uncertainty of the panel of experts.

3) ROC Analysis: Finally, we perform ROC analysis to evaluate the ability to discriminate between the least uncertain and most uncertain cases, based on the predictive entropy. To do so, we threshold the entropy of the panel of experts based on the median value, unless stated otherwise.

V. Results

Table I reports the results for the Gleason2019 experiment, including the best performing model from [10], based on Annotator Confusion Estimation from [12] which aims to model the reliability of each expert in a setting of noisy labels. The unweighted Cohen's kappa values between the set of annotators are also included [22], [23]. We observe that all methods perform on par with the best results of [10] in terms of Gleason grading performance, although the different methods did not outperform the majority vote baseline.

More importantly, when evaluating the task of modeling the reference-standard label distribution, all methods demonstrate KL divergence values slightly lower than the majority vote

TABLE I
TARGET TASK PERFORMANCE ON THE GLEASON2019 GRADING DATASET

Model	Target Task			Uncertainty Estimation		
	Tumor v. Benign (Acc.)	High v. Low (Acc.)	Kappa	$D_{KL} \downarrow$	Spearman	AUC
Majority Vote	$91.65_{\pm 2.03}$	$79.02_{\pm 3.07}$	$0.571_{\pm .061}$	$0.524_{\pm .031}$	$0.189_{\pm .067}$	$61.39_{\pm 3.54}$
Histogram of Doctors	$91.60_{\pm 1.92}$	$78.54_{\pm 3.30}$	$0.565_{\pm .054}$	$0.470_{\pm .028}$	$0.243_{\pm .061}$	$64.98_{\pm 2.53}$
Ensemble (1 CNN/doc)	$\mathbf{91.72_{\pm 1.89}}$	$78.83_{\pm 3.02}$	$0.572_{\pm .057}$	$0.471_{\pm .024}$	$0.247_{\pm .048}$	$65.18_{\pm 2.15}$
LDM	$91.63_{\pm 1.77}$	$79.08_{\pm 2.88}$	$0.572_{\pm .056}$	$0.462_{\pm .027}$	$0.251_{\pm .055}$	$65.41_{\pm 2.67}$
CLDM	$91.63_{\pm 1.87}$	$\mathbf{79.20_{\pm 2.90}}$	$\mathbf{0.574_{\pm .059}}$	$\mathbf{0.461_{\pm .028}}$	$\mathbf{0.253_{\pm .051}}$	$\mathbf{65.53_{\pm 2.57}}$
ACE [10]	92	80	-	-	-	-
Pathologists [22], [23]	-	-	0.40 - 0.60	-	-	-

Mean and standard deviation values are calculated across three repetitions. KL-divergence (D_{KL}) measures similarities, and spearman's coefficients and AUC values measure correlation, between ground truth and predicted distributions for each method.

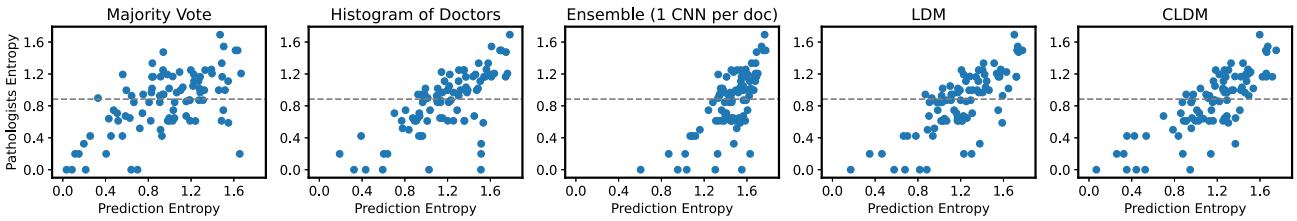

Fig. 3. Scatter plots, visualizing correlation between the entropy of the predictive distribution and the entropy of the ground-truth label distribution.

baseline (0.46 - 0.47 vs. 0.52). In other words, the predicted distributions are more similar to the ground-truth label distribution, with the CLDM demonstrating the highest similarity. A similar performance gap is found when evaluating the Spearman's correlation between the entropy values of the prediction and the ground-truth distribution, and the AUC value for discriminating between cases above or below the mean entropy value of the panel of pathologists. With a median ground-truth entropy of zero, we decided to set the threshold at the mean value instead. However, both these metrics only show moderate levels of correlation between the entropy of the predictive distribution and the ground truth distribution (Spearman's values up to 0.25). Note, although variations in performance within each fold were small, larger variations beween the folds resulted in the relatively high standard deviation values.

Table II reports the main results for the ISUP grading experiments: both the dense and the simulated limited label scenario. We start by looking at the grading performance of each method in both scenarios, reported by the Cohen's kappa coefficients. The histogram of doctors method reports the highest kappa scores for the grading task in both settings. We observe that training on the noisy majority vote label, based on only five observers in the limited label scenario, results in a significant drop in ISUP grading performance. Here, training on the full label distribution (HoD) helps to mitigates issues related to label noise in the limited label scenario. We see a similar performance gap when comparing classification accuracy in the limited label scenario between the majority vote baseline and the HoD method with accuracy values of 54.81 and 64.07 respectively. In both settings, differences in classification performance between the different methods show similar trends compared to the Cohen's kappa coefficients. However, in the dense label scenario the majority vote method reaches the highest classification accuracy of 69.26, followed by the HoD method with an accuracy of 67.41.

Afterwards, we evaluate the more important task of modeling the full ground-truth distribution of labels. Regarding the similarities between the ground-truth label distribution and the predictive distributions, measured by the KL divergence term, we observe that our proposed method demonstrates the highest similarity. Furthermore, the LDM and CLDM seem the least affected by limiting the amount of annotations in the training set, with similar KL divergence values in both scenarios and a larger performance gap with other methods in the limited label scenario.

Fig. 3 plots the entropy of the predictive distribution against the entropy of the 20 pathologists, for each point in the test set in the dense label scenario. Here, the median entropy of the panel of pathologists is highlighted for reference. We observe that the LDM and the CLDM show the highest correlation, with Spearman's coefficients up to 0.728 for the dense label scenario (Table II). Furthermore, when measuring the ability to discriminate between WSIs above and below the median entropy of the panel of experts, using ROC analysis based on the entropy of the prediction, the LDM and CLDM demonstrate the highest AUC values (81.10 - 86.42). Again, the performance differences with the other methods seem highest under the limited label scenario. To verify the statistical significance of the observed performance differences, we calculate the p-value using a permutation test based on 10 k samples. Limited in statistical power, by the size of the dataset, we don't observe a significant difference in performance by the LDM with all baseline methods. However, all methods significantly outperform the majority vote method on all uncertainty estimation metrics ($p < 0.05$), except for the AUC value of the ensemble in the dense label scenario.

Fig. 4 shows predictions for three random test cases. We observe that the majority vote method often under estimates uncertainty, with predictions resembling one-hot vectors. In contrast, the other methods show more "soft" predictions. When looking at a random set of individual predictions, represented by the lighter colored histograms, the deep ensembles method demonstrate the most diverse set of predictions compared to the LDM and the CLDM.

TABLE II
TARGET TASK PERFORMANCE ON THE ISUP GRADING DATASET, FOR BOTH DENSE AND LIMITED LABEL SCENARIOS

Model	Dense Label Scenario				Limited Label Scenario ($D = 5$)			
	Target Task	Uncertainty Estimation			Target Task	Uncertainty Estimation		
	Kappa	$D_{KL} \downarrow$	Spearman	AUC	Kappa	$D_{KL} \downarrow$	Spearman	AUC
Majority Vote	$0.815_{\pm.011}$	$0.417_{\pm.013}$	$0.540_{\pm.038}$	$74.52_{\pm1.98}$	$0.672_{\pm.025}$	$0.530_{\pm.033}$	$0.434_{\pm.021}$	$68.23_{\pm1.17}$
Histogram of Doctors	$\mathbf{0.819_{\pm.008}}$	$0.257_{\pm.003}$	$0.671_{\pm.007}$	$82.04_{\pm1.48}$	$\mathbf{0.788_{\pm.014}}$	$0.326_{\pm.010}$	$0.595_{\pm.020}$	$79.52_{\pm.816}$
Ensemble (1 CNN/doc)	$0.811_{\pm.006}$	$0.342_{\pm.001}$	$0.649_{\pm.020}$	$80.71_{\pm.934}$	$0.745_{\pm.004}$	$0.464_{\pm.004}$	$0.608_{\pm.006}$	$77.74_{\pm.450}$
LDM	$0.799_{\pm.014}$	$\mathbf{0.256_{\pm.007}}$	$\mathbf{0.728_{\pm.021}}$	$85.38_{\pm1.11}$	$0.768_{\pm.008}$	$\mathbf{0.297_{\pm.011}}$	$\mathbf{0.653_{\pm.026}}$	$\mathbf{82.83_{\pm2.17}}$
CLDM	$0.797_{\pm.025}$	$0.263_{\pm.014}$	$0.727_{\pm.040}$	$\mathbf{86.42_{\pm2.75}}$	$0.762_{\pm.010}$	$0.300_{\pm.013}$	$0.632_{\pm.036}$	$81.10_{\pm1.86}$

Mean and standard deviation values are reported for a total of fifteen independent train runs. KL-divergence (D_{KL}) measures similarities, and spearman's coefficients and AUC values measure correlation, between reference-standard and predicted distributions for each method.

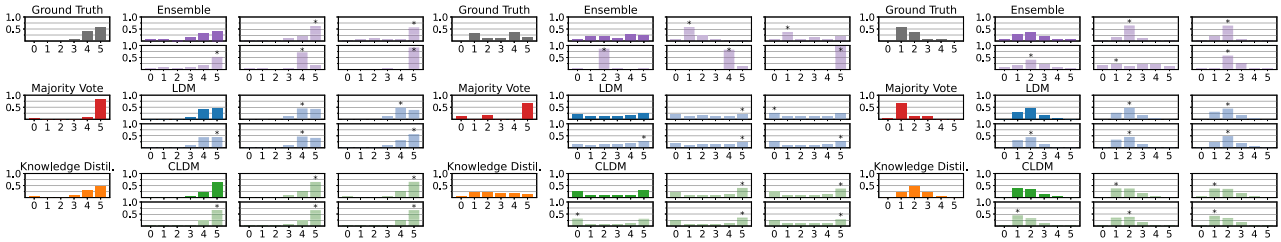

Fig. 4. Predictive distributions for three random cases from the ISUP grading test set, to illustrate the similarities with the ground truth distribution. For methods involving a distribution of predictions (the ensemble, LDM and CLDM), the mean prediction is shown (dark color), as well as five random individual predictions (lighter color). For each individual prediction, the predicted class (argmax) is highlighted with an asterisk.

To analyze the influence of $\mathbf{z}$ on the classifier $p_\theta(y|\mathbf{x}, \mathbf{z})$, we also evaluate a LDM trained with a latent space of two dimensions, see Fig. 5. By taking different samples of $\mathbf{z}$, corresponding to different positions in the 2D latent space, we observe the impact of the latent vector on the predicted class labels. Here, each color corresponds to a different class label. We observe that the resulting distribution of predicted class labels closely resembles the distribution of ground-truth annotations, demonstrating the ability of a single trained model to capture a diverse and realistic distribution of labels.

Finally, we evaluate the ability to filter out difficult cases at various uncertainty thresholds, and compute the grading performance on the remaining test set (see Fig. 6). By removing the top τ percent of cases, with $\tau \in [0, 1]$, we observe an improved grading performance for all methods. We observe that the deep ensembles and HoD methods benefit most by removing only a small subset of the test set (with $\tau < 0.2$). Combined with the uncertainty estimation results from Table II, this could suggest an optimal workflow where ambiguous cases are flagged by a LDM, followed by ISUP grading using a superior classifier like an ensemble.

The main findings from the previous two experiments are corroborated in the results on the task of tumor budding classification, reported in Table III. See Fig. 7 for predictive distributions for three random samples from the test set. For consistency, we report Cohen's Kappa coefficients in Table III, however, we see similar performance differences when evaluating classification accuracy. Specifically, the CLDM achieves the highest accuracy of 82.54, closely followed by the other methods with the majority vote baseline reaching an accuracy of 82.03. We observe that all methods outperform the majority vote method on all uncertainty related metrics. These results demonstrate that training on the full label distribution instead of the majority vote, seems imperative to improve the similarities between the ground-truth and the predicted distribution, without impacting the performance on the classification task. When evaluating the Spearman's correlation coefficient, we observe moderate levels of correlation between the predictive entropy and the entropy of the panel of pathologists (0.476 - 0.556). Similarly, ROC analysis based on the median entropy of annotators results in moderate AUC values (75.47 - 78.36). However, by performing ROC analysis to evaluate the ability to detect cases for which no majority vote was found among at least three out of the seven annotations, we observe high AUC values (93.1 - 94.3) for all methods trained using the full label distribution. With AUC values of: 78.61, 93.23, 94.27, 93.08, 94.17, corresponding to the same order of methods reported in Table III. In contrast to the previous two experiments, the LDM does not outperform the other methods. Instead, the deep ensemble, where each member is trained on the labels from one of the nine pathologists, outperforms the other methods slightly on all metrics except classification accuracy.

Finally, we include a performance sensitivity analysis for the LDM using the following hyper-parameters: the size of the

TABLE III
TARGET TASK PERFORMANCE ON THE BUDDING DATASET

Model	Target Task	Uncertainty Estimation		
	Kappa	$D_{KL} \downarrow$	Spearman	AUC
Maj. Vote	$0.647_{\pm.001}$	$0.392_{\pm.027}$	$0.476_{\pm.016}$	$75.47_{\pm1.18}$
HoD	$0.645_{\pm.006}$	$0.190_{\pm.003}$	$0.527_{\pm.007}$	$76.83_{\pm.633}$
Ensemble	$0.646_{\pm.003}$	$\mathbf{0.187_{\pm.000}}$	$\mathbf{0.556_{\pm.004}}$	$\mathbf{78.36_{\pm.031}}$
LDM	$0.643_{\pm.005}$	$0.198_{\pm.003}$	$0.531_{\pm.009}$	$77.30_{\pm.367}$
CLDM	$\mathbf{0.654_{\pm.003}}$	$0.196_{\pm.001}$	$0.538_{\pm.003}$	$77.49_{\pm.290}$

Mean and Standard deviation values are calculated across three repetitions. KL-divergence (D_{KL}) measures similarities, and spearman's coefficients and AUC values measure correlation between ground truth and predicted distributions for each method.

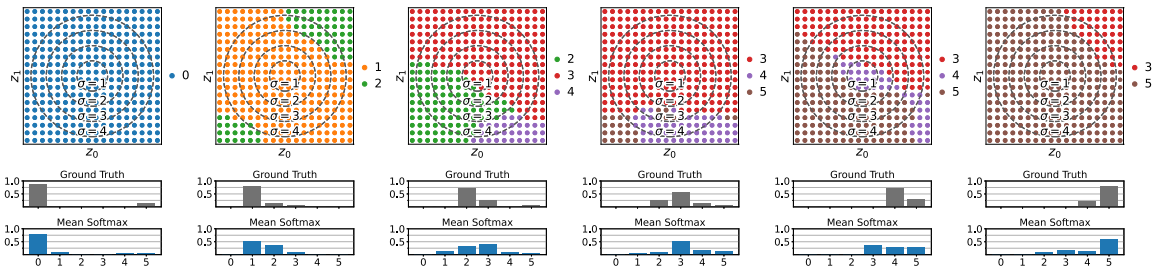

Fig. 5. Visualization of the latent space (LDM), and its influence on classification results for the ISUP grading task. Evaluating the classifier $p(y|\mathbf{x}, \mathbf{z})$, using 18×18 samples $\mathbf{z}$ from the 2D latent space for six random WSIs (one per ISUP grade) from the test set. Contours corresponding to the distance, in standard deviations, to the origin for the standard Gaussian distribution are included. Colors represent the predicted class (argmax) for each forward pass, demonstrating a more diverse set of predictions for cases where annotators disagree more.

TABLE IV
SENSITIVITY ANALYSIS WHEN REPEATING THE ISUP GRADING EXPERIMENT FOR THE DENSE LABEL SCENARIO USING DIFFERENT HYPERPARAMETER SETTING

Model	Target Task	Uncertainty Estimation		
	Kappa	$D_{KL} \downarrow$	Spearman	AUC
LDM default params	$0.799_{\pm.014}$	$0.256_{\pm.007}$	$0.728_{\pm.021}$	$85.38_{\pm1.11}$
LDM ($\lvert\mathcal{S}_n^d\rvert = 1$)	$0.793_{\pm.011}$	$0.305_{\pm.055}$	$0.603_{\pm.163}$	$80.03_{\pm9.93}$
LDM ($\lvert\mathcal{S}_n^d\rvert = 4$)	$0.812_{\pm.008}$	$0.273_{\pm.009}$	$0.711_{\pm.022}$	$88.48_{\pm.429}$
LDM ($\lvert\mathcal{S}_n^d\rvert = 32$)	$0.794_{\pm.022}$	$0.271_{\pm.002}$	$0.712_{\pm.027}$	$87.95_{\pm1.58}$
LDM ($K = 1$)	$0.795_{\pm.015}$	$0.264_{\pm.008}$	$0.698_{\pm.016}$	$87.15_{\pm.591}$
LDM ($K = 4$)	$0.797_{\pm.009}$	$0.256_{\pm.009}$	$0.739_{\pm.018}$	$\mathbf{89.01}_{\pm.440}$
LDM ($K = 32$)	$0.803_{\pm.010}$	$0.257_{\pm.008}$	$0.729_{\pm.022}$	$88.21_{\pm.447}$
LDM ($K = 128$)	$0.805_{\pm.009}$	$\mathbf{0.256}_{\pm.008}$	$0.729_{\pm.020}$	$88.45_{\pm.287}$
LDM ($dim(z) = 1$)	$0.825_{\pm.019}$	$0.267_{\pm.002}$	$0.716_{\pm.024}$	$88.59_{\pm1.99}$
LDM ($dim(z) = 2$)	$\mathbf{0.842}_{\pm.017}$	$0.261_{\pm.007}$	$0.720_{\pm.017}$	$86.53_{\pm2.25}$
LDM ($dim(z) = 32$)	$0.810_{\pm.022}$	$0.264_{\pm.011}$	$0.719_{\pm.033}$	$88.08_{\pm1.17}$
LDM ($dim(z) = 64$)	$0.810_{\pm.015}$	$0.259_{\pm.011}$	$\mathbf{0.743}_{\pm.032}$	$88.37_{\pm1.41}$
ResNet50 Backbone				
Maj. Vote	$0.725_{\pm.016}$	$0.476_{\pm.013}$	$0.568_{\pm.048}$	$74.61_{\pm4.63}$
HoD	$\mathbf{0.732}_{\pm.006}$	$\mathbf{0.344}_{\pm.010}$	$0.629_{\pm.016}$	$82.87_{\pm1.34}$
Ensemble	$0.719_{\pm.015}$	$0.379_{\pm.002}$	$\mathbf{0.702}_{\pm.008}$	$\mathbf{86.45}_{\pm.338}$
LDM	$0.723_{\pm.006}$	$0.353_{\pm.003}$	$0.641_{\pm.006}$	$83.94_{\pm.245}$
CLDM	$0.723_{\pm.012}$	$0.359_{\pm.008}$	$0.638_{\pm.007}$	$81.94_{\pm.803}$

Mean and standard deviation values are calculated across three repetitions. KL-divergence (D_{KL}) measures similarities, and spearman's coefficients and auc values measure correlation, between ground truth and predicted distributions for each method. We included the previously reported results using the default hyperparameters, as well the results when repeating the experiment using a different (ResNet50) backbone model.

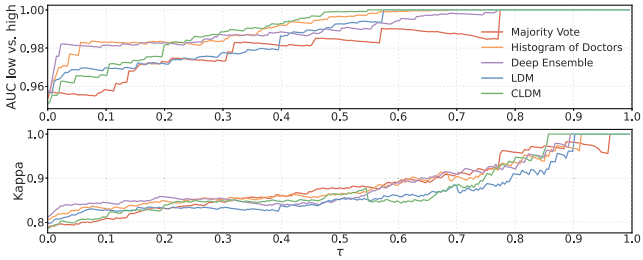

Fig. 6. ISUP grading performance at various thresholds, by removing the top τ percent of cases with the highest predictive entropy, and evaluate on the remaining test set. Each method is evaluated using ROC analysis for the task of separating low (grade 0-2) from high ISUP grades (3-5) (top), and the Cohen's kappa coefficients (bottom) at each threshold.

support set, the amount of samples taken during inference K, and the dimension of the latent space. We do so by changing one parameter at a time and repeating the ISUP grading experiment for the dense label scenario. Table IV reports the results. We observe the biggest drop in performance when using a support set of size one, or with k=1. Explained by the limited information captured within the latent space in these scenarios. As the number of latent dimensions increases, the quality of the uncertainty estimates appears to improve. However, most improvements fall within each other's standard deviation ranges. Furthermore, the most significant performance gain compared to baseline models is already achieved by the LDM based on a single latent dimension. Overall these results demonstrate robustness to small changes within the most important set of hyper-parameters. Lastly, we repeat the ISUP grading experiment using a different backbone model $h(\mathbf{x})$, based on a ResNet50 architecture. We observe that the ISUP grading performance (0.719 - 0.732) degrades significantly compared to the reported results based on the DenseNet backbone (up to 0.819). Although the target task performance is not the main focus, these results indicate that the different methods may not have been trained properly. Further optimization might necessitate a distinct set of hyperparameters for the alternative image backbone, which is considered beyond the scope of the current work. This could apply to either the training of the pre-trained backbone model or each specific downstream method.

VI. DISCUSSION AND CONCLUSION

Ambiguities are prevalent in many medical imaging tasks, and even experienced doctors demonstrate high-inter observer variability for the most challenging cases [3], [5]. We aim to model the full label distribution of a panel of experts, to capture the disagreement between a set of experts.

Throughout extensive experiments on three datasets, we show that we can increase the similarities between the predictive and the ground-truth label distributions by training on the full label distribution instead of the majority vote label. More importantly, we demonstrate that the predictive uncertainty can be used as a surrogate value for the inter-observer variability of a panel of experts by correlating entropy values between both distributions. Furthermore, we show that the different classifiers demonstrate improved tumor grading performance when removing the most uncertain cases in the ISUP grading task. These results suggest that the predictive uncertainty of these methods can be used to

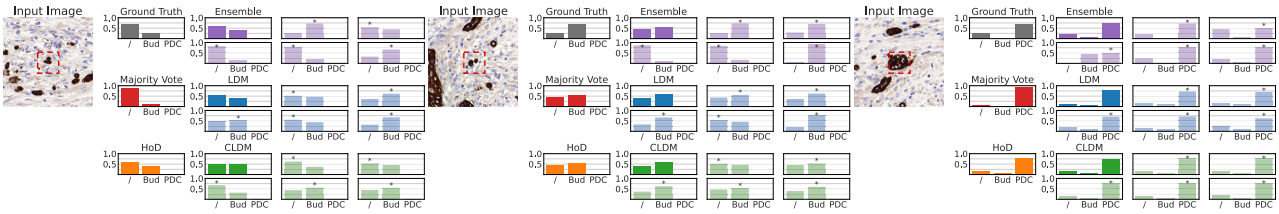

Fig. 7. Predictive distributions for three random cases from the budding test set, to illustrate the similarities with the ground truth distribution. For methods involving a distribution of predictions (the ensemble, LDM and CLDM), the mean prediction is shown (dark color), as well as five random individual predictions (lighter color). For each individual prediction, the predicted class (argmax) is highlighted with an astrix.

flag complex cases for further analysis in an automated clinical diagnosis workflow.

We observe that the proposed LDM outperforms the other methods on two prostate tumor grading datasets regarding the metrics used to evaluate the similarities between the ground-truth label distribution and the predictive distribution. The performance gap is most significant when evaluating ISUP grading on a whole-slide image level in both the dense and the limited label scenarios. Here, we observe that the predictive uncertainty of the LDM correlates the most with the entropy of the label distribution determined by 20 pathologists.

A similar performance gap between the LDM and the majority vote method is found on the tumor budding dataset. However, the LDM and the CLDM are outperformed by the deep ensembles method based on nine independently trained networks. Here, we note that the budding task is different in nature due to the labels being categorical: the object is either a bud, a PDC or something different, instead of ordinal such as ISUP grading (with increasing tumor grades labelled zero through 5). These results indicate that the LDM performs especially well in capturing annotator variability on ordinal classification tasks, such as grading. Future work is required to confirm these findings on other datasets, with varying annotation densities in different limited label scenarios.

Throughout the experiments, conditioning the latent space on the input image (CLDM) only slightly impacts performance compared to the unconditioned model (LDM), either positively or negatively, depending on the experiment. These results are in line with the equivalent probabilistic U-Net [15], which showed negligible differences between the conditional and unconditional prior latent space. The current settings for the DeepSet encoder showed to work well throughout all experiments and the performance sensitivity analysis demonstrated the LDM to be robust towards small changes.

When comparing the diversity of predictions, the ensemble demonstrates the highest diversity compared to the (C)LDM. However, when analyzing the 2D latent space of the LDM, we saw that the LDM leads to a diverse and realistic set of classification predictions. Furthermore, ensembles require significantly more computational resources to train and evaluate one network per doctor in the training set, which does not scale well to larger datasets like the ISUP grading dataset. The total of 300 individual ensemble members required a computational budget of around 50 GPU hours. In contrast, although the individual training cycle of the LDM takes longer, the total computational budget required summed up to just 30 hours. During inference, differences in computational requirements are negligible and well under a minute per patient. Here, the LDM is able to capture annotator variability in a single model and can even outperform the deep ensembles method in uncertainty estimation. Furthermore, it could potentially give valuable insight into annotator behaviour by analyzing structures in the continuous latent space. The related probabilistic U-Net demonstrates more consistent improvements on deep ensembles in [15]. However, they compare with a regularly trained ensemble based on random initializations [19]. For a fair comparison, we train the ensemble to model the full label distribution by training on individual annotators.

Although the different methods performed on par with previously reported results in classification accuracy on the Gleason grading experiment [10], correlations between the predictive uncertainty and entropy of the ground-truth label distributions are lower compared to the other experiments. This could be due to only moderate levels of agreement between the annotators, with Cohen's kappa coefficients between 0.40 and 0.60 [22], [23], making this task especially difficult.

Interestingly, the histogram of doctors baseline shows consistent improvements on the majority vote method on both classification accuracy and the ability to model the full label distribution. By simply using soft targets during training, the HoD approach is a cheap alternative to obtain robust estimates of uncertainty without a trade-off in classification performance. However, this approach will not provide more insights into individual annotator behavior, such as the individual outputs by the LDM and the deep ensembles.

All in all, this work demonstrates that the LDM and related methods are able to model a distribution of labels. We demonstrated that the predictive uncertainty is informative for the accuracy of the prediction and the disagreement between a panel of experts. We believe that future work in medical imaging should benefit from all available annotations.

Acknowledgment

The authors would like to thank Knut and Alice Wallenberg foundation is acknowledged for generous support.

References

[1] J. I. Epstein, "An update of the Gleason grading system," *J. Urol.*, vol. 183, no. 2, pp. 433–440, 2010.
[2] W. C. Allsbrook Jr et al., "Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists," *Hum. Pathol.*, vol. 32, no. 1, pp. 74–80, 2001.
[3] R. N. Flach et al., "Significant inter- and intralaboratory variation in Gleason grading of prostate cancer: A nationwide study of 35,258 patients in The Netherlands," *Cancers*, vol. 13, no. 21, 2021, Art. no. 5378.
[4] L. Egevad et al., "Utility of pathology imagebase for standardisation of prostate cancer grading," *Histopathology*, vol. 73, no. 1, pp. 8–18, 2018.
[5] L. Egevad et al., "Standardization of Gleason grading among 337 European pathologists," *Histopathology*, vol. 62, no. 2, pp. 247–256, 2013.
[6] W. Bulten et al., "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study," *Lancet Oncol.*, vol. 21, pp. 233–241, Feb. 2020.

[7] W. Bulten et al., "Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists," *Modern Pathol.*, vol. 34, no. 3, pp. 660–671, 2021.
[8] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Proc. Adv. Neural Inf. Process*, 2015, pp. 3483–3491.
[9] M. Zaheer et al., "Deep sets," in *Proc. Adv. Neural Inf. Process.*, 2017, pp. 3391–3401.
[10] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, 2020, Art. no. 101759.
[11] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021.
[12] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proc. IEEE Comp. Vis. Pattern Recognit.*, 2019, pp. 11244–11253.
[13] H. Zhu, J. Shi, and J. Wu, "Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation," in *Proc. Int. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 576–584.
[14] H. Le, D. Samaras, T. Kurc, R. Gupta, K. Shroyer, and J. Saltz, "Pancreatic cancer detection in whole slide images using noisy label annotations," in *Proc. Int. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 541–549.
[15] S. A. A. Kohl et al., "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process*, Montréal, Canada, 2018, pp. 6965–6975.
[16] C. F. Baumgartner et al., "PHiSeg: Capturing uncertainty in medical image segmentation," in *Proc. Int. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 119–127.
[17] S. A. Kohl et al., "A hierarchical probabilistic U-Net for modeling multi-scale ambiguities," 2019, *arXiv.1905.13077*.
[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.
[19] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process*, 2017, pp. 6402–6413.
[20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014.
[21] I. Higgins et al., "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl
[22] G. Nir et al., "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts," *Med. Image Anal.*, vol. 50, pp. 167–180, 2018.
[23] G. Nir et al., "Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images," *JAMA Netw. Open*, vol. 2, no. 3, pp. e190442–e190442, 2019.
[24] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.
[25] J. I. Epstein et al., "A contemporary prostate cancer grading system: A validated alternative to the Gleason score," *Eur. Urol.*, vol. 69, no. 3, pp. 428–435, 2016.
[26] W. Bulten et al., "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge," *Nature Med.*, vol. 28, no. 1, pp. 1–10, 2022.
[27] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Comp. Vis. Pattern Recognit*, 2017, pp. 2261–2269.
[28] J. Linmans, J. van der Laak, and G. Litjens, "Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks," in *Proc. Med. Img. Deep Learn.*, 2020, pp. 465–478.
[29] J. Bokhorst et al., "Assessment of individual tumor buds using keratin immunohistochemistry: Moderate interobserver agreement suggests a role for machine learning," *Modern Pathol.*, vol. 33, no. 5, pp. 825–833, 2020.
[30] G. Hinton et al., "Distilling the knowledge in a neural network," *NIPS Depp Learn. Workshop*, 2014.
[31] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.
[32] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.